

Mathematics of Image and Data Analysis

Math 5467

Lecture 26: Momentum Descent

Instructor: Jeff Calder

Email: jcalder@umn.edu

<http://www-users.math.umn.edu/~jwcalder/5467S21>

Announcements

- Project 3 due May 9.
- Final exam Thurs/Fri (will be available online Wed evening)
 - Exam will have 2 questions. Will be easier than typical homework questions.
 - You must work on the exam by yourself.
 - You can use the class notes and homework.
 - You cannot look up solutions online, or post the questions on Math.StackExchange

• Office Hours: Tuesday May 4 @ 9am-10am
Last time

- Gradient descent

Today

- Momentum descent

Gradient Descent

Gradient descent is one of the most important algorithms in many areas of science and engineering. To minimize an objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, gradient descent iterates

$$(1) \quad x_{k+1} = x_k - \alpha \nabla f(x_k)$$

until convergence. The parameter $\alpha > 0$ is the time step (often called the *learning rate* when using gradient descent to train machine learning algorithms).

Assumptions on f

We assume the objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a smooth function that admits a global minimizer $x_* \in \mathbb{R}^n$. That is

$$f(x_*) \leq f(x)$$

for all $x \in \mathbb{R}^n$. We denote the optimal value of f by $f_* := f(x_*)$.

Sublinear convergence rate

We say ∇f is *L-Lipschitz continuous* if

$$(2) \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n.$$

Theorem 1. *Assume ∇f is L-Lipschitz and that $\alpha \leq \frac{1}{L}$. Then for any integer $t \geq 1$ we have*

$$(3) \quad \min_{0 \leq k \leq t} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_*)}{\alpha t}.$$

Remark 2. The theorem says, with very few assumptions on f , that gradient descent converges at a rate of $O\left(\frac{1}{t}\right)$ to a critical point of f , in the sense that $\nabla f \sim \frac{1}{t} \rightarrow 0$. Since f is not assumed to be convex, critical points need not be minimizers and could also include saddle points.

Linear convergence

To obtain a better convergence rate, we need to make an additional assumption about how flat f can be at minima. We say that f is μ -strongly convex if

$$(4) \quad f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|x - y\|^2$$

for all $x, y \in \mathbb{R}^n$.

Note: If we take $x = x_*$ then $\nabla f(x_*) = 0$ and we get

$$(5) \quad f(y) \geq f_* + \frac{\mu}{2}\|y - x_*\|^2.$$

Polyak-Lojasiewicz (PL) inequality

If f is μ -strongly convex, then f satisfies the PL inequality

$$(6) \quad \frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(x) - f_*)$$

for all $x \in \mathbb{R}^n$.

Remark 3. The PL inequality is weaker than strong convexity, and even nonconvex functions can satisfy it (as an exercise, show that $f(x) = x^2 + 3 \sin^2(x)$ satisfies the PL inequality (6) with $\mu = \frac{1}{32}$, but f is not convex).

Linear convergence

Theorem 4. Assume f satisfies the PL inequality (6), ∇f is L -Lipschitz, and take $\alpha \leq \frac{1}{L}$. Then for any integer $t \geq 0$ we have

$$(7) \quad f(x_t) - f_* \leq (1 - \alpha\mu)^t (f(x_0) - f_*).$$

Remark 5. The best rate is obtained by taking $\alpha = \frac{1}{L}$ in which case we obtain

$$f(x_t) - f_* \leq \left(1 - \frac{\mu}{L}\right)^t (f(x_0) - f_*).$$

The ratio $\kappa = \frac{\mu}{L}$ is called the *condition number* of f (or rather of $\nabla^2 f$), and controls the rate of convergence of gradient descent.

Convergence of minimizers

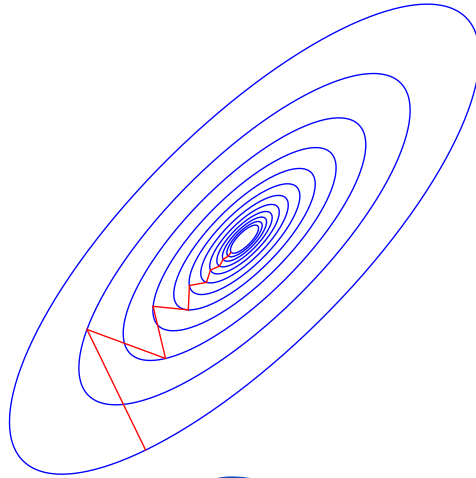
Remark 6. It is also natural to ask how quickly x_k is converging to x_* . For this, we require strong convexity. If f is μ -strongly convex then we have

$$\frac{\mu}{2} \|x_t - x_*\|^2 \leq f(x_t) - f_* \leq (1 - \alpha\mu)^t (f(x_0) - f_*).$$

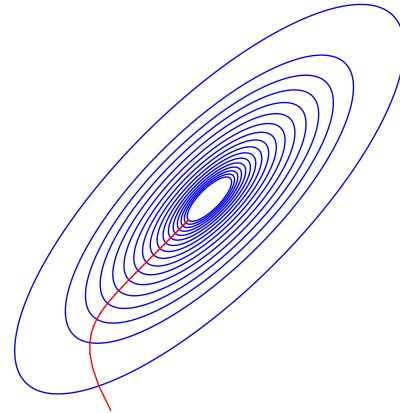
Ex:

Issues with gradient descent

$$f(x, y)$$



(a) $\alpha = 0.05$, 12 steps



(b) $\alpha = 0.01$, 50 steps

Figure 1: Gradient descent on a parabolic function with different choices of time steps. For larger time steps the iterations bounce back and forth, limiting progress towards the minimizer, while for smaller time steps the descent path is more direct.

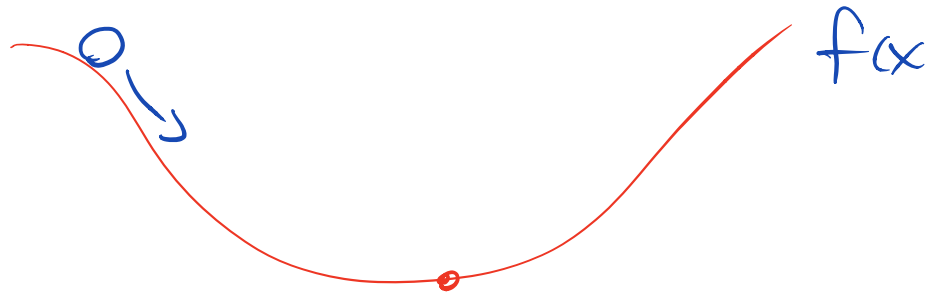
Momentum descent

One of the oldest momentum based methods is the heavy ball method of Polyak. The heavy ball method iterates *G.D.* *momentum*

$$(8) \quad x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}),$$

where α is the time step and $\beta \in [0, 1]$ is the momentum parameter, where $x_1 = x_0$.

- The idea is that the descent direction has *memory*, or *momentum*. This averages out the bouncing effect in gradient descent, and accelerates convergence when the descent directions align over many iterations (near the minimizer).
- As we will see, the descent equations share similarities with the equations of motion for a ball rolling down the energy landscape, so it is also called the *heavy ball method*.



Heavy ball method

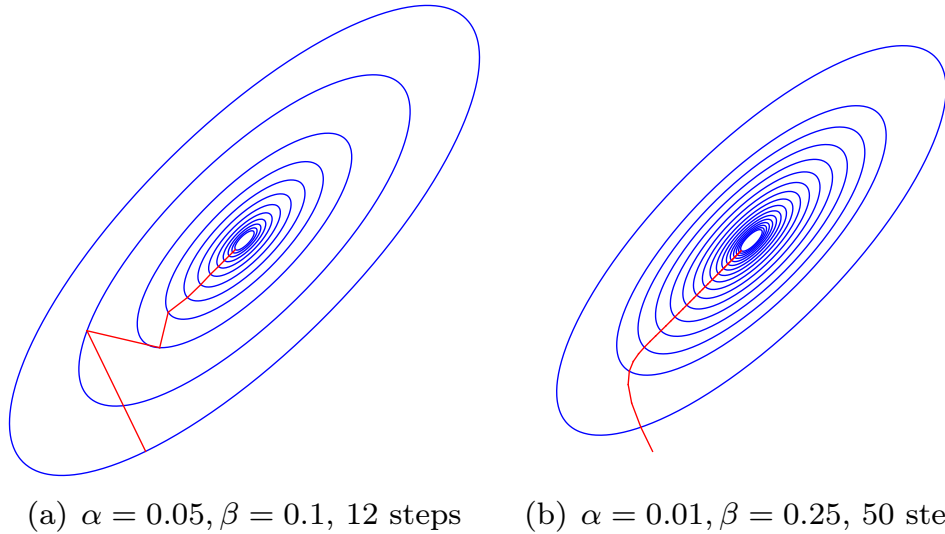
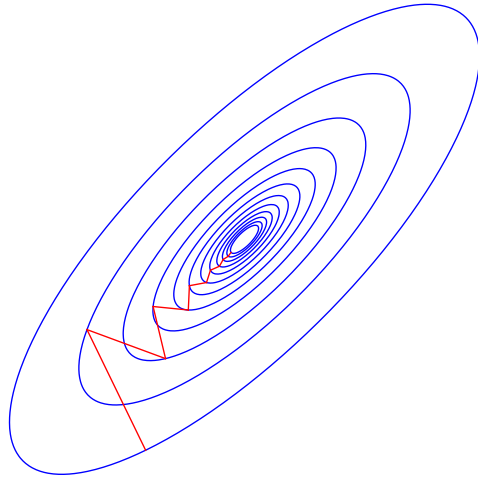
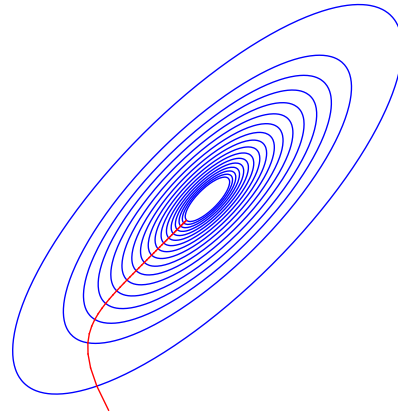


Figure 2: Heavy ball method for different choices of time step and momentum parameter. Momentum acts to average out the descent direction in time, limiting the bouncing effect for larger time steps. Momentum builds up speed and makes more progress towards the minimizer in the same number of steps as gradient descent.

Recall: Gradient descent



(a) $\alpha = 0.05$, 12 steps



(b) $\alpha = 0.01$, 50 steps

Figure 3: Gradient descent on a parabolic function with different choices of time steps. For larger time steps the iterations bounce back and forth, limiting progress towards the minimizer, while for smaller time steps the descent path is more direct.

Continuum perspective: Gradient Descent

For gradient descent (1), we can rewrite the equation as

$$x_{k+1} = x_k - \alpha \nabla f$$

$$x'(\alpha t) \approx \frac{x_{k+1} - x_k}{\alpha} = -\nabla f(x_k).$$

By assuming $x_k = x(\alpha t)$ for a smooth curve $x(t)$, we find that the left hand side is merely a forward differences approximation for $x'(t)$, and so gradient descent is equivalent in the continuum to the ordinary differential equation (ODE)

$$x'(t) = -\nabla f(x(t)).$$

Continuum perspective: Heavy ball method

On the other hand, when we rearrange the heavy ball method iteration (8) in a similar way, we obtain

$$(9) \quad \frac{x_{k+1} - 2x_k + x_{k-1}}{\alpha} + \frac{1 - \beta}{\alpha}(x_k - x_{k-1}) = -\nabla f(x_k).$$

Heavy Ball

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$

$$x_{k+1} - x_k + \beta(x_{k-1} - x_k) = -\alpha \nabla f(x_k)$$

$$x_{k+1} - x_k + x_{k-1} - x_k - (x_{k-1} - x_k) + \beta(x_{k-1} - x_k) = -\alpha \nabla f(x_k)$$

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{\alpha} + \frac{(1-\beta)(x_k - x_{k-1})}{\alpha} = -\nabla f(x_k)$$

Discrete approximations of derivatives

Exercise 7. Show that

$$\frac{x(t) - x(t-h)}{h} = x'(t) + O(h),$$

$\sim \leq C|$

and

$$\frac{x(t+h) - 2x(t) + x(t-h)}{h^2} = x''(t) + O(h^2)$$

for a smooth curve $x(t)$. To do this, use the Taylor expansions

$$x(t \pm h) = x(t) \pm x'(t)h + \frac{h^2}{2}x''(t) \pm \frac{h^3}{6}x'''(t) + O(h^4).$$

\triangle

Continuum perspective: Heavy ball method

Recalling the heavy ball method can be written as x'' x' $x_k = x(\sqrt{\alpha} k)$

$$x'' \rightarrow \frac{x_{k+1} - 2x_k + x_{k-1}}{\alpha} + \frac{1 - \beta}{\sqrt{\alpha}} \left(\frac{x_k - x_{k-1}}{\sqrt{\alpha}} \right) = -\nabla f(x_k).$$

we can use Exercise 7 to see that this is a discretization of the ODE

$$(10) \quad x''(t) + \frac{1 - \beta}{\sqrt{\alpha}} x'(t) = -\nabla f(x(t)). \quad ma = F$$

- These are the equations of motion (Newton's law) for the motion of an object under the force $-\nabla f(x(t))$ with damping/friction coefficient $\frac{1 - \beta}{\sqrt{\alpha}}$.
- This suggests choosing β so that

$$\frac{1 - \beta}{\sqrt{\alpha}} = c$$

that is $\beta = 1 - c\sqrt{\alpha}$.

Analysis of heavy ball method

The analysis of the heavy ball method is more involved, compared to gradient descent. We will analyze the method in the special case of solving the linear system

$$(11) \quad Ax = b,$$

where A is an $n \times n$ positive definite and symmetric matrix (e.g., a discrete Laplacian). We can solve this equation by minimizing

$$f(x) = \frac{1}{2}x^T Ax - x^T b.$$

Note that $\nabla f(x) = Ax - b$.

$$f(x) = \frac{1}{2}x^T A x - x^T b.$$

Gradient descent

Theorem 8. Suppose x_k satisfies

$$(12) \quad x_{k+1} = x_k - \alpha(Ax_k - b)$$

for all $k \geq 1$, and assume $\alpha \leq \frac{1}{L}$. Then we have

$$(13) \quad (1 - \alpha L)^k \leq \frac{\|x_k - x_*\|}{\|x_0 - x_*\|} \leq (1 - \alpha \mu)^k.$$

The proof is left as an exercise (or look in class notes).

Heavy ball method

Theorem 9. Suppose x_k satisfies

$$(14) \quad x_{k+1} = x_k - \alpha(Ax_k - b) + \beta(x_k - x_{k-1})$$

for all $k \geq 2$ and $x_1 = x_0$. Let $\alpha \leq \frac{1}{L}$ and assume

$$(15) \quad (1 - \sqrt{\alpha\mu})^2 \leq \beta \leq 1.$$

Then for all $k \geq 2$ we have

$$(16) \quad \|x_k - x_*\|^2 + \|x_{k+1} - x_*\|^2 \leq 2\beta^k \|x_0 - x_*\|^2.$$

Remark 10. Theorem 9 suggests that the optimal choice for β is $\beta = (1 - \sqrt{\alpha\mu})^2$. If we also take $\alpha = \frac{1}{L}$ and write $\kappa = \frac{\mu}{L}$ then

$$\text{Heavy ball: } \|x_k - x_*\| \leq \sqrt{2}(1 - \sqrt{\kappa})^k \|x_0 - x_*\|,$$

compared to

$$\text{Gradient Descent: } \|x_k - x_*\| \leq (1 - \kappa)^k \|x_0 - x_*\|.$$

Nesterov Acceleration

$$\text{Friction} = \frac{1-\beta}{\sqrt{\alpha}}$$

For convex functions that may not be strongly convex, Nesterov's accelerated gradient method can accelerate convergence. Nesterov's method iterates

$$x_{k+1} = y_k - \alpha \nabla f(y_k), \quad y_{k+1} = x_{k+1} + \frac{k-1}{k+2} (x_{k+1} - x_k).$$

$\beta_k \rightarrow 1$
Vanishing friction

- The first step is gradient descent and the second step is a momentum correction, but the friction decreases to zero over time.
- Nesterov proved the method converges at an $O\left(\frac{1}{t^2}\right)$ rate for convex functions, which is a substantial improvement over the $O\left(\frac{1}{t}\right)$ rate for gradient descent.
- It turns out the $O\left(\frac{1}{t^2}\right)$ rate is provably optimal for minimizing convex functions with first order methods.

Proof of Theorem 9

$$(*) \quad x_{k+1} = x_k - \alpha (Ax_k - b) + \beta (x_k - x_{k-1})$$

Let v_1, v_2, \dots, v_n be orthonormal eigenvectors of A , $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ be eigenvalues

$$\mu = \lambda_1, \quad L = \lambda_n.$$

Take dot product of $(*)$ with v_i

$$x_{k+1}^T v_i = x_k^T v_i - \alpha ((Ax_k)^T v_i - b^T v_i) + \beta (x_k^T v_i - x_{k-1}^T v_i)$$

$$\text{Let } c_k = x_k^T v_i, \quad b_i = \alpha b^T v_i$$

$$c_{k+1} = c_k - \alpha \left(\underbrace{x_k^T A v_i}_{= \lambda_i v_i} - \alpha^T b_i \right) + \beta (c_k - c_{k-1})$$

$$= c_k - \alpha (\lambda_i c_k - \alpha^T b_i) + \beta (c_k - c_{k-1})$$

$$= (1 + \beta - \alpha \lambda_i) c_k - \beta c_{k-1} + b_i$$

$$\begin{bmatrix} c_{k+1} \\ c_k \end{bmatrix} = \begin{bmatrix} 1 + \beta - \alpha \lambda_i & -\beta \\ 1 & 0 \end{bmatrix} \begin{bmatrix} c_k \\ c_{k-1} \end{bmatrix} + \begin{bmatrix} b_i \\ 0 \end{bmatrix}$$

Write $B = \begin{bmatrix} 1 + \beta - \alpha \lambda_i & -\beta \\ 1 & 0 \end{bmatrix}$

$$\begin{bmatrix} c_{k+1} \\ c_k \end{bmatrix} = B \begin{bmatrix} c_k \\ c_{k-1} \end{bmatrix} + \begin{bmatrix} b_i \\ 0 \end{bmatrix}$$

Iterate k times:

$$\begin{bmatrix} c_{k+1} \\ c_k \end{bmatrix} = B^k \begin{bmatrix} c_1 \\ c_0 \end{bmatrix} + \sum_{j=0}^{k-1} B^j \begin{bmatrix} b_i \\ 0 \end{bmatrix}$$

Need eigenvalues of B :

$$B = \begin{bmatrix} 1 + \beta - \alpha \lambda_i & -\beta \\ 1 & 0 \end{bmatrix}$$

Characteristic polynomial

$$p_i(\lambda) = \det(\lambda I - B)$$

$$= \det \begin{bmatrix} \lambda - 1 - \beta + \alpha \lambda_i & \beta \\ -1 & \lambda \end{bmatrix}$$

$$= (\lambda - 1 - \beta + \alpha \lambda_i) \lambda + \beta$$

$$= \lambda^2 - (1 + \beta - \alpha \lambda_i) \lambda + \beta$$

Eigenvalues satisfy $p_i(\lambda) = 0$,

$$\lambda = \frac{1}{2} \left(1 + \beta - \alpha \lambda_i \pm \sqrt{(1 + \beta - \alpha \lambda_i)^2 - 4\beta} \right)$$

discriminant.

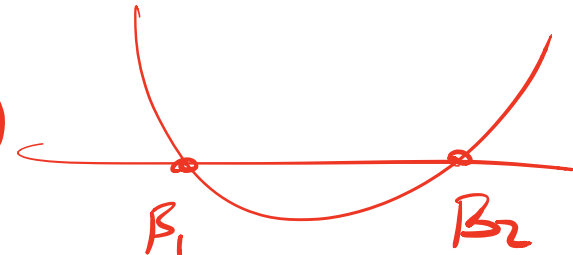
Discriminant is ≤ 0 when

$$(1 + \beta - \alpha \lambda_i)^2 - 4\beta \leq 0$$

$$\beta^2 + 2\beta(1 - \alpha d_i) + (1 - \alpha d_i)^2 - 4\beta \leq 0$$

$$\beta^2 + 2\beta(1 - \alpha d_i - 2) + (1 - \alpha d_i)^2 \leq 0$$

Roots are $= -(1 + \alpha d_i)$



$$\beta = \frac{1}{2} \left(2(1 + \alpha d_i) \pm \sqrt{4(1 + \alpha d_i)^2 - 4(1 - \alpha d_i)^2} \right)$$

$$= 1 + \alpha d_i \pm \sqrt{\cancel{1} + 2\alpha d_i + \cancel{\alpha^2 d_i^2} - \cancel{1} + 2\alpha d_i - \cancel{\alpha^2 d_i^2}}$$

$$= 1 + \alpha \lambda_i \pm \sqrt{4\alpha \lambda_i}$$

$$\begin{aligned}\beta &\geq 1 + \alpha \lambda_i - 2\sqrt{\alpha \lambda_i} \\ &= (1 - \sqrt{\alpha \lambda_i})^2\end{aligned}$$

When $\beta \geq (1 - \sqrt{\alpha \lambda_i})^2$, then the discriminant is ≤ 0 and eigenvalues λ are complex-valued. In this case

$$\lambda = \frac{1}{2} \left(\underbrace{1 + \beta - \alpha \lambda_i}_{\text{real part}} \pm \underbrace{\sqrt{(1 + \beta - \alpha \lambda_i)^2 - 4\beta}}_{\text{imaginary part}} \right)$$

$$|\lambda|^2 = \frac{(1 + \beta - \alpha \lambda_i)^2}{4} + \frac{4\beta - (1 + \beta - \alpha \lambda_i)^2}{4}$$

$$= \frac{4\beta}{4} = \beta$$

So $|\lambda| = \sqrt{\beta}$ when $\beta \geq (1 - \sqrt{\mu})^2$

This implies $\|Bx\| \leq \sqrt{\beta} \|x\|$

$$\begin{bmatrix} c_{k+1} \\ c_k \end{bmatrix} = B^k \begin{bmatrix} c_1 \\ c_0 \end{bmatrix} + \sum_{j=0}^{k-1} B^j \begin{bmatrix} b_i \\ 0 \end{bmatrix}$$

Use geometric series

$$(*) \quad \sum_{j=0}^{k-1} B^j = (I - B)^{-1} - B^k (I - B)^{-1}$$

$$(**) \quad (I - B)^{-1} \begin{bmatrix} b_i \\ 0 \end{bmatrix} = (I - B^k)^{-1} \begin{bmatrix} x_{\dagger}^T v_i \\ x_{\dagger}^T v_i \end{bmatrix}$$

when

$$Ax_* = b.$$

$$\begin{bmatrix} c_{k+1} \\ c_k \end{bmatrix} = B^k \begin{bmatrix} c_1 \\ c_0 \end{bmatrix} + (I - B)^{-1} \begin{bmatrix} b_i \\ 0 \end{bmatrix}$$

$$\leftarrow -B^k (I - B)^{-1} \begin{bmatrix} b_i \\ 0 \end{bmatrix}$$

$$= B^k \begin{bmatrix} c_1 \\ c_0 \end{bmatrix} + \begin{bmatrix} x_*^T v_i \\ x_*^T v_i \end{bmatrix} - B^k \begin{bmatrix} x_*^T v_i \\ x_*^T v_i \end{bmatrix}$$

$$- \cancel{B^k (I - B)^{-1} \begin{bmatrix} b_i \\ 0 \end{bmatrix}}$$

$$\left\| \begin{bmatrix} x_{k+1}^T v_i - x_*^T v_i \\ x_k^T v_i - x_*^T v_i \end{bmatrix} \right\|^2 = \left\| B^k \begin{bmatrix} c_1 - x_*^T v_i \\ c_0 - x_*^T v_i \end{bmatrix} \right\|^2$$

$$\text{Apply } \|B^k x\| = (\sqrt{\beta})^k \|x\| \dots$$

